Estadística multivariada, 1 sem. 2019

Juan Carlos Castillo & Alejandro Plaza

Sesión 2: Bases

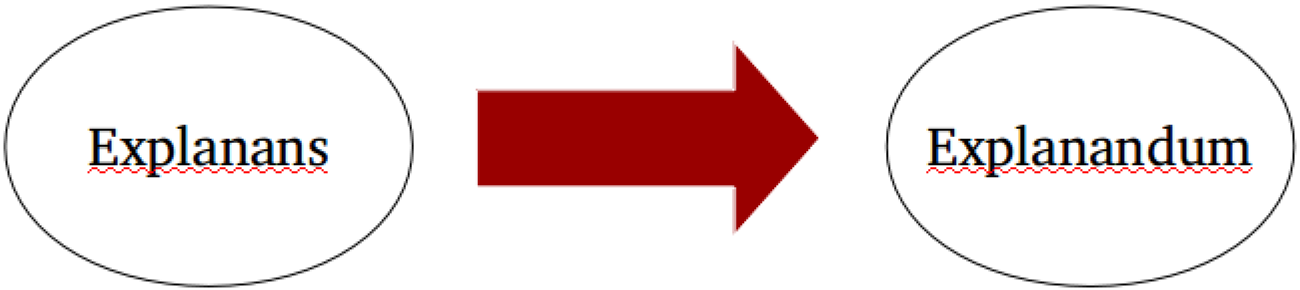
Contenidos

1. Repaso de sesión anterior
2. Datos
3. Variables
4. Bases Estadística descriptiva: Tendencia Central y Variabilidad
5. Prueba de Hipótesis
6. Correlación

1. Repaso sesión anterior

La explicación en ciencias sociales

El concepto de explicación en ciencias sociales

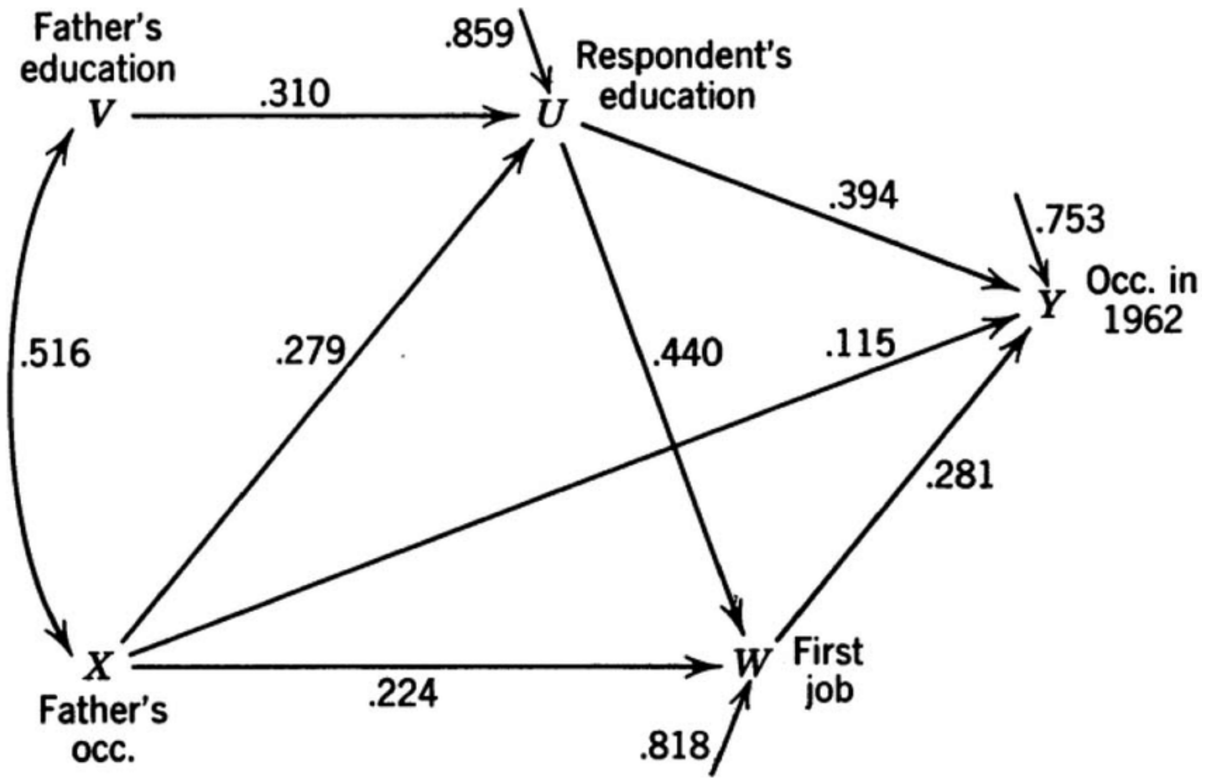


- Explanandum: el fenómeno que predentemos explicar (precisión, relevancia y variabilidad).
- Explanans: lo que genera la aparición del fenómeno (lógica, eficacia y claridad.)

Modalidades de explicación en ciencias sociales (Linares, 2018)

- Por leyes de cobertura.
- Explicación funcional.
- Explicación Estadística.
- Explicación "como si".
- *Explicación por mecanismos*.

...Volviendo a Pedro, Juan & Diego



2. Datos

Datos y su representación

- Los datos miden al menos una *característica* de a los menos una *unidad* en a lo menos *un punto en el tiempo*
    - Ejemplo: La tasa de natalidad en Chile el 2017 fue de 1,8 hijos (por mil habitantes)
    - Característica (variable) : Tasa de natalidad
    - Unidad: País
    - Punto en el tiempo: 2017

Base de Datos

- Los datos se almacenan en una estructura de base de datos
- Base de datos:
  - cada fila representa una unidad o caso (ej: un entrevistado)
  - cada columna una variable (ej: edad)

m0_sexo	m0_edad	m01
Mujer	64	Educacion Basica o Preparatoria incompleta
Mujer	60	Educacion Media o Humanidades incompleta
Mujer	26	Educacion Media o Humanidades incompleta
Hombre	51	Universitaria incompleta
Hombre	69	Educacion Media o Humanidades completa
Hombre	62	Educacion Media o Humanidades incompleta
Hombre	36	Educacion Basica o Preparatoria completa
Mujer	54	Educacion Media o Humanidades completa
Mujer	32	Tecnica Superior incompleta
Hombre	38	Educacion Media o Humanidades completa

Ejemplos

1. Encuesta Centro de Estudios Públicos
2. Encuesta CASEN
3. Encuesta Lapop

3. Variables

Definición

Una variable representa cualquier cosa o propiedad que varia y a la cuál se le asigna un valor. Es decir:

$$Variable \neq Constante$$

Pueden ser visibles o no visibles (latentes). Y además se pueden agrupar en:

- Variables discretas (Rango finito de valores):
  - Dicotómicas
  - Politómicas
- Variables continuas.
  - Rango (teóricamente) infinito de valores.

Escalas de medición de variables

Escalas (Stevens, 1946): la asignación de medición se manifiesta en distintos niveles o escalas. (acrónimo clave: **NOIR**)

Table 2.1 Levels of Measurement

Scale Type	Defining Characteristic	Properties of Numbers	Examples
Nominal	Numbers are used instead of words.	Identity or equality	SS#s; football players' jersey numbers; numerical codes for nonquantitative variables, such as sex or psychiatric diagnoses
Ordinal	Numbers are used to order a hierarchical series.	Identity + rank order	Ranking of athletes or teams; percentile scores
Interval	Equal intervals between units but no true zero.	Identity + rank order + equality of units	Fahrenheit and Celsius temperature scales; calendar time
Ratio	Zero means "none of" whatever is measured; all arithmetical operations possible *and* meaningful.	Identity + rank order + equality of units + additivity	Measures of length; periods of time

Escalas de Variables

Escala	Ejemplos	Rango	Differencia	Cero
Nominal	Género, nacionalidad	No	No	No
Ordinal	Nivel educacional	Sí	No	No
Intervalar	Prestigio ocupacional	Sí	Sí	No
Ratio	Ingreso, número de niños	Sí	Sí	Sí

Tipos de datos en relación a escalas de medición.

- Datos categóricos: pueden ser medidos sólo mediante escalas nominales, u ordinales en caso de orden de rango
- Datos continuos:
    - Medidos en escalas intervalares o de razón
    - Pueden ser transformados a datos categóricos

Tipos de análisis en relación a tipos de datos.

	Categórica	Continua	Categórica(y)/Categórica(x)	Contínua(y)/Categórica(x)
Ejemplo	Estatus Ocupacional	Ingreso	Estatus Ocupacional (Y) / Género (X)	Ingreso (Y) / Género (X)
Tabla	Sin problemas	Necesidad de recodificar	Tabla de Contingencia	Clasificar Y
Gráfico	Barras	Histograma / boxplot	Gráfico de barras condicionado	Histograma, box plot condicionado
Estadística	Frecuencias, proporciones, odds	Media, medidas de dispersión.	Proporciones condicionadas, odds condicionados	Media condicionada, Mediana condicionada

Tipos de análisis estadístico bivariado.

Variable independiente x	Variable dependiente Categórica	Variable dependiente Continua
Categórica	Análisis de tabla de Contigencia, Chi2	Análisis de Varianza ANOVA, Prueba T
Continua	Regresión Logística (y probit)	Regresión Lineal

4. Bases estadística descriptiva:

Medidas de tendencia central y variabilidad

Tendencia Central

- Moda: valor que ocurre más frecuentemente
- Mediana: valor medio de la distribución ordenada. Si N es par, entonces es el promedio de los valores medios
- Media o promedio aritmético: suma de los valores dividido por el total de casos
    - Desventaja: influencia de valores extremos

Dispersión: Rangos

- Rango: distancia entre los puntos extremos de la distribución

- Rango intercuartil / semi intercuartil
    - Intercuartil: rango de acumulación del 50% de los datos Ej: 77,5-55=22,5
    - Semi- intercuartil: la mitad 22,5/2=11,25

Cuartiles	**Percentiles**	**Puntajes**
1	25	32,5
2	50	55
3	75	77,5
4	100	100

Dispersión: Varianza

- Suma de las diferencias al cuadrado de cada valor (x) y el promedio de la distribución divididos por el total menos 1. Formalmente:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$$

- Considerando N-1 para la varianza de la muestra.

ID	**Pje (x)**	$x - \bar{x}$	$(x - \bar{x})^2$
1	6	0.4	0.16
2	4	-1.6	2.56
3	7	1.4	1.96
4	2	-3.6	12.96
5	9	3.4	11.56
Sum	28	0	29.2
Prom	5.6		

$$\sigma^2 = \frac{(29.2)}{5-1}$$

$$= 7.3$$

Desviación Estándar

- Raiz Cuadrada de la varianza.
  - Se interpreta como la variabilidad promedio de los puntajes desde un punto de referencia común: el promedio de los datos.
  - Expresada en la mismas unidades que los puntajes.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

En el ejemplo anterior:

$$\sigma = \sqrt{\frac{(29.2)^2}{5-1}}$$

$$= 2.7$$

5. Prueba de Hipótesis

Hipótesis

- Proposición respecto a uno o varios parámetros
- Prueba de hipótesis
    - Determinar si la hipótesis es congruente con los datos obtenidos en la muestra
    - Por ejemplo, si mi hipótesis es que hombres y mujeres poseen diferente rendimiento en matemática, el objetivo del análisis es encontrar diferencias estadísticamente significativas entre ambos grupos en la muestra

Prueba de Hipótesis y significación estadística

- Las hipótesis no pueden ser aceptadas o descartadas 100% a partir de los estadígrafos
- El rechazo de hipótesis tiene que ver con el concepto de PROBABILIDAD
  *Ej: ¿con qué nivel de probabilidad puedo decir que existen diferencias entre hombres y mujeres en rendimiento en matemáticas?
- Por lo tanto, el elemento central en la prueba de hipótesis es establecer es la probabilidad de error que estamos cometiendo en la inferencia

Prueba de Hipótesis y significación estadística

- Dada la probabilidad asociada a la inferencia, es imposible demostrar que algo es verdadero.
- Para hacer frente a esta situación, se establecen dos tipos de hipótesis:

  -Hipótesis nula (H_0): no existen diferencias -Hipótesis alternativa (H_a): existen diferencias
- Objetivo de la investigación: rechazar H_0

Ejemplo

¿Tiene el entrenamiento en matemáticas un impacto en mayor puntaje SIMCE?

$$H_0 : \mu_0 = \mu_1 \vee \mu_{entren} = \mu_{pob}$$

$$H_a : \mu_0 > \mu_1 \vee \mu_{entren} > \mu_{pob}$$

Tipos posibles de error

Rechazar H_0 cuando esta es verdadera (Error tipo I o α)

No rechazar H_a cuando esta es falsa (Error tipo I o β)

7. Correlación

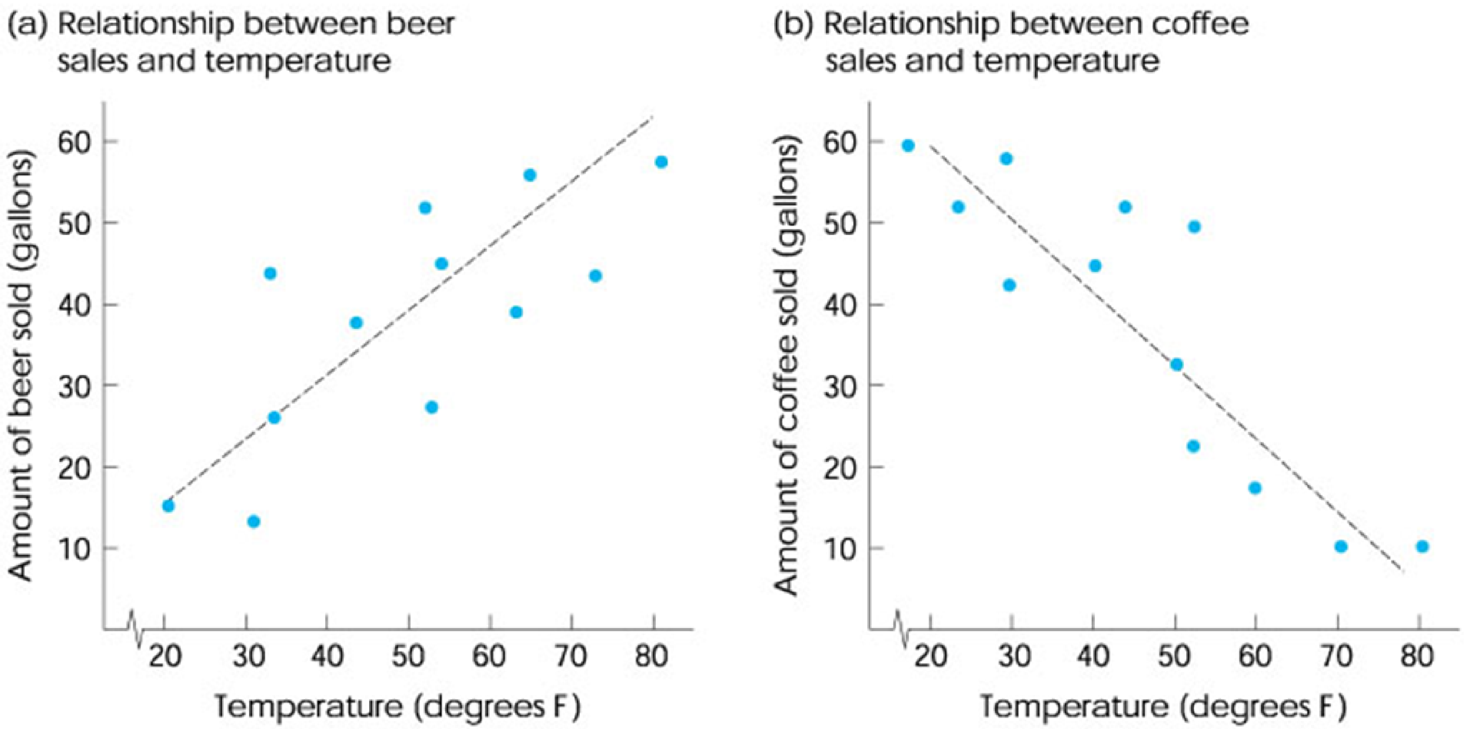
Bases correlación

- Es una técnica estadística usada para medir y describir la relación entre dos variables numéricas (nivel de medición de intervalo o de razón)
- La medida más común de correlación es el coeficiente de correlación de Pearson (r).
- Da cuenta de: Intensidad de la asociación y dirección
- Su rango de variación es entre **-1 y 1**

Dirección

1. Correlación Positiva: cuando dos variables se mueven en la misma dirección. En otras palabras, cuando valores altos de una variable están asociados a valores altos en otra variable (años de educación e ingreso)
2. Correlación negativa: cuando las dos variables se mueven en direcciones opuestas Valores altos de una variable están asociadas con valores bajos de la otra (nivel de eficacia colectiva vecinal y sensación de inseguridad)

Correlación: Positiva y Negativa



Forma e intesidad de relación

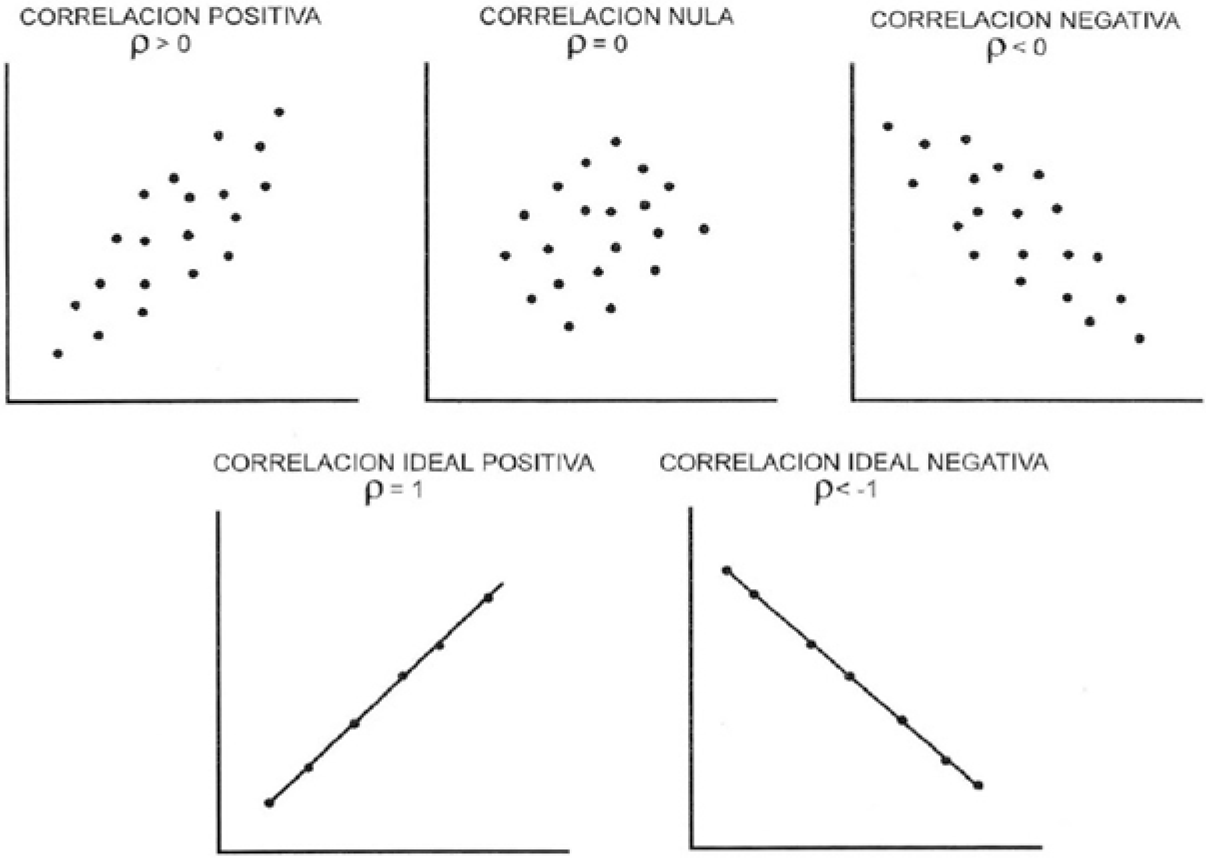
Asociación lineal: cuando los puntos en un diagrama tienden a tener forma de una línea recta

Correlación de Pearson mide cuán bien los puntos en un gráfico se ajustan a una relación lineal

Grado de intensidad ¿Cuán exactamente se ajustan los datos a la forma lineal específica? El grado de intensidad es medido por el valor numérico de los valores del coeficiente de correlación r

- Entre -1.0 y +1.0
- Correlación r =0 indica ausencia absoluta de relación lineal
- Correlación -1.0 indica correlación lineal perfecta negativa
- Correlación +1.0 indica correlación lineal perfecta positiva

Nubes de puntos y correlaciones



Correlación de Pearson

Mide el grado y la dirección de una relación lineal entre dos variables (de nivel de medición intervalo/razón)

$$r = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

Para calcular la correlación necesitamos algo que llamaremos suma de productos de las desviaciones (SP) de X e Y

$$SP = \sum (x - \bar{x})(y - \bar{y})$$

Esto es análogo a la suma de cuadrados (SC), solo que ahora se mide la covarabilidad (COVARIANZA) entre dos variables en vez de la variación de una sola variable.

$$SC = \sum (x - \bar{x})$$

Correlación de Pearson

La suma de productos (SP) se usa para calcular el coeficiente de correlación Pearson r junto con la suma de cuadrados de X y de Y

$$r = \frac{SP(xy)}{\sqrt{SC_x SC_y}}$$

o bien

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

Aspectos a considerar

Correlación NO implica causalidad: x no es causa de y ni y es causa de x; solo están asociados.

La correlación debería estar informada por teoría que haga inteligible la asociación entre X e Y.

Que no exista correlación lineal no significa (necesariamente) que las variables no estén asociadas de otra forma (curvilínea, por ejemplo)

Aspectos a considerar: Ejemplo

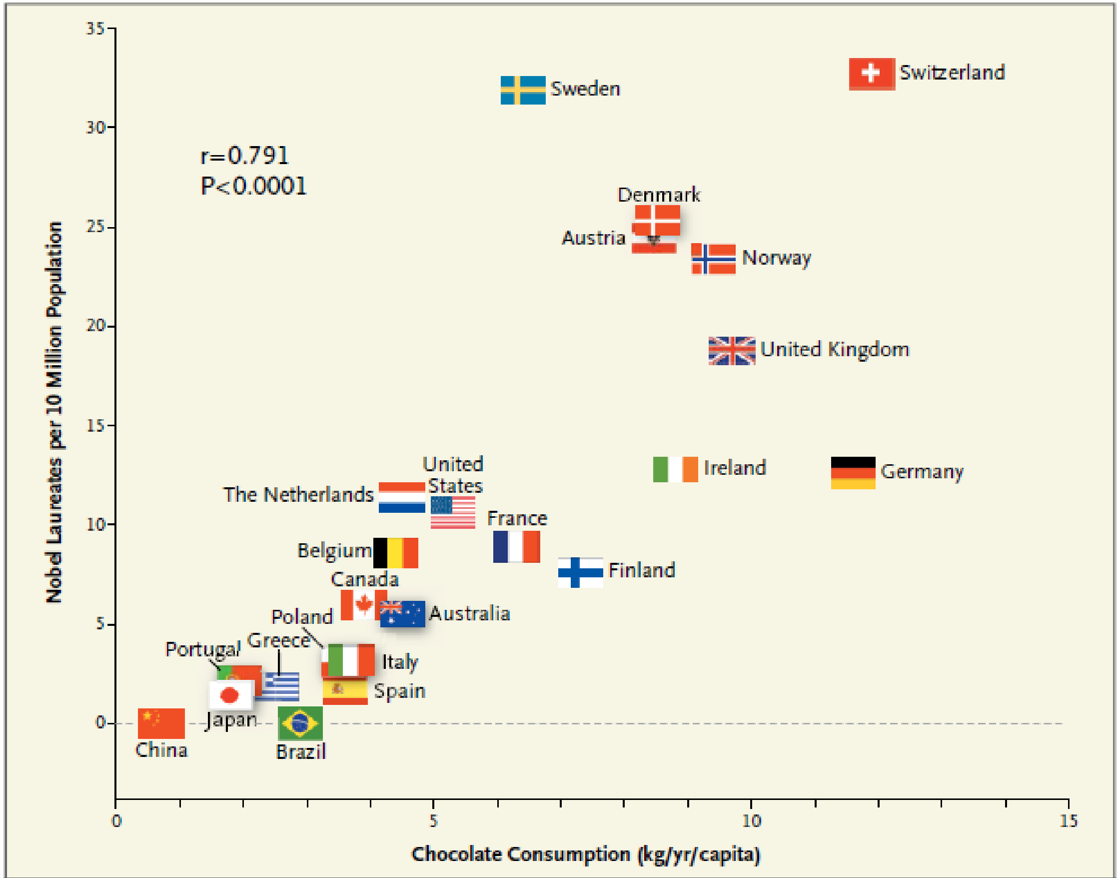


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Ejemplo de correlación

Estimar la correlación entre puntaje en lenguaje (x) y puntaje en matemáticas (y):

id	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})*(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
1	17	24	-3	3	-9	9	9
2	19	23	-1	2	-2	1	4
3	14	22	-6	1	-6	36	1
4	22	17	2	-4	-8	4	16
5	15	23	-5	2	-10	25	4
6	26	21	6	0	0	36	0
7	23	18	3	-3	-9	9	9
8	21	17	1	-4	-4	1	16
9	28	21	8	0	0	64	0
10	15	24	-5	3	-15	25	9
Sum					-63	210	68
Prom	20	21					

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}$$

$$= \frac{-63}{\sqrt{210*68}}$$

$$= -0.5272$$

Ejemplo cálculo en R

- 1.Ingreso manual de datos

```
x <- c(17,    19,    14,    22,    15,
       26,    23,    21,    28,    15)
y <- c(24,    23,    22,    17,    23,
       21,    18,    17,    21,    24)
```

- 2.Promedios

```
prom_x=mean(x)
prom_x
```

```
## [1] 20
```

```
prom_y=mean(y)
prom_y
```

```
## [1] 21
```

Ejemplo cálculo en R

- 3.Numerador de Pearson: suma de productos de diferencias del Promedio

```
prod_difs_xy <- (x-(mean(x)))*(y-(mean(y)))
sum_prod_difs_xy <- sum(prod_difs_xy)
sum_prod_difs_xy
```

```
## [1] -63
```

Ejemplo cálculo en R

- 4.Denominador Pearson: Raiz del producto de la suma de cuadrados de x por la de y

```
dif_x2<- (x-(mean(x)))^2
sum_dif_x2 <- sum(dif_x2)
sum_dif_x2
```

```
## [1] 210
```

```
dif_y2<- (y-(mean(y)))^2
sum_dif_y2 <- sum(dif_y2)
sum_dif_y2
```

```
## [1] 68
```

Ejemplo cálculo en R

- 5.Pearson

```
corr=sum_prod_difs_xy/sqrt((sum_dif_x2)*(sum_dif_y2))
corr
```

```
## [1] -0.5272013
```

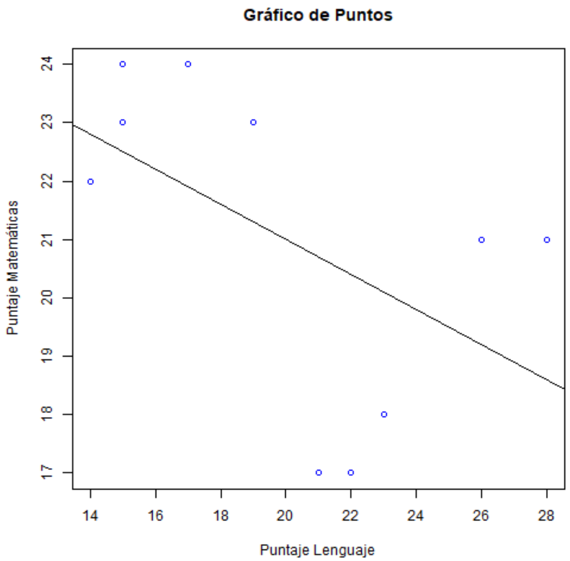
… y por comando en R

```
cor(x,y)
```

```
## [1] -0.5272013
```

Demostración en R

```
plot(x, y,  col = "blue", main = "Gráfico de Puntos",
xlab = "Puntaje Lenguaje", ylab = "Puntaje Matemáticas")
abline(lm(y ~ x))
```



Ejercicio práctico

¿Cuál es la relación entre la temperatura y las ventas de helado?

A partir de la siguiente tabla calcule la correlación (y covarianza) entre la temperatura y las ventas de helado.

Temperatura	**Ventas de Helado**
66	8
72	11
77	15
84	20
83	21
71	11
65	8
70	10